

INVESTIGATIONS OF THE NON-LINEAR LMC CEPHEID PERIOD-LUMINOSITY RELATION WITH TESTIMATOR AND SCHWARZ INFORMATION CRITERION METHODS

S. M. KANBUR

Department of Physics, State University of New York at Oswego, Oswego, NY 13126

C. NGEOW

Department of Astronomy, University of Illinois, Urbana, IL 61801

A. NANTHAKUMAR AND R. STEVENS

Department of Mathematics, State University of New York at Oswego, Oswego, NY 13126

Draft version February 1, 2008

ABSTRACT

In this paper, we investigate the linearity versus non-linearity of the Large Magellanic Cloud (LMC) Cepheid period-luminosity (P-L) relation using two statistical approaches not previously applied to this problem: the estimator method and the Schwarz Information Criterion (SIC). The estimator method is extended to multiple stages for the first time, shown to be unbiased and the variance of the estimated slope can be proved to be smaller than the standard slope estimated from linear regression theory. The Schwarz Information Criterion (also known as the Bayesian Information Criterion) is more conservative than the Akaike Information Criterion and tends to choose lower order models. By using simulated data sets, we verify that these statistical techniques can be used to detect intrinsically linear and/or non-linear P-L relations. These methods are then applied to independent LMC Cepheid data sets from the OGLE project and the MACHO project, respectively. Our results imply that there is a change of slope in longer period ranges for all of the data sets. This strongly supports previous results, obtained from independent statistical tests, that the observed LMC P-L relation is non-linear with a break period at/around 10 days.

Subject headings: Cepheids — distance scale — Stars: fundamental parameters — methods: statistical

1. INTRODUCTION

The cornerstone of the extra-galactic distance scale is the Cepheid Period-Luminosity (P-L) relation defined by the Large Magellanic Cloud (LMC) Cepheids. The assumed linear relation of the LMC Cepheid P-L relation, which is linear in $\log(P)$, with P the pulsation period in days, has been under debate due to recent results that this relation could be non-linear (Tammann & Reindl 2002; Kanbur & Ngeow 2004; Sandage et al. 2004; Ngeow et al. 2005). These authors contended that the existing Cepheid data in the LMC strongly suggested the LMC P-L relation is consistent with two lines of significantly differing slopes with a break at/around a period of 10 days. This is referred as the non-linearity of the Cepheid P-L relation in this paper. Arguments for choosing the fiducial period at 10 days can be found in Kanbur & Ngeow (2004), Sandage et al. (2004), Ngeow et al. (2005) and Ngeow & Kanbur (2006a). Furthermore, Kanbur & Ngeow (2004, 2006), Sandage et al. (2004), Ngeow et al. (2005) and Ngeow & Kanbur (2006c) examined various factors that may cause the non-linearity of the LMC P-L relation, including the observing strategies, photometric errors, extinction errors, removal of outliers, sample selection, number of long period Cepheids in the samples and contamination of overtone Cepheids. They found that none of these remedies or any combination of them could be responsible for the observed non-linear LMC P-L relation. As argued in Ngeow & Kanbur (2006c), rigorous statistical tests are

needed to test the linearity versus the non-linearity of the LMC P-L relation.

In our previous studies, the F -test (e.g. Weisberg 1980) has been applied to the OGLE (Optical Gravitational Lensing Experiment) and MACHO (MAssive Compact Halo Objects project) Cepheid data, in Kanbur & Ngeow (2004) and Ngeow et al. (2005) respectively, to test for the non-linearity of the LMC P-L relation. In such a formulation, the full and reduced models are models with four and two parameters respectively. This test looks at the change in the mean residual sum of squares between the full and reduced model divided by the mean residual sum of squares in the full model (see equation [5] of Kanbur & Ngeow 2004). This test statistic can be formulated as the difference in slopes between short and long period slopes divided by the standard error of that difference. Hence if the number and nature of the long/short period data are such that the long/short period slope is estimated with a large error, then the F -value will be low and return a non-significant result. Thus the F -test is sensitive to the number of data points on either side of the period cut at 10 days. The OGLE and MACHO data sets we used in Kanbur & Ngeow (2004) and Ngeow et al. (2005), respectively, do have adequate number of long and short period Cepheids for the application of the F -test. The F -test has returned a significant result when testing the non-linearity of the P-L relation in both of the data sets.

Nevertheless, the results that suggesting a non-linear LMC P-L relation are still controversial. As we empha-

size that statistical tests are needed, however claims of linear LMC P-L relation in the literature lack of rigorous statistical tests. In this paper, we apply two additional statistical tests, the estimator method and the Schwarz Information Criterion method, to examine the non-linearity of the LMC Cepheid P-L relation. These tests will be complementary to the F -test carried out in previous studies since they will serve to check and verify the results obtained from the F -test. In this way previous conclusions about the non-linear LMC P-L relation are considerably strengthened. Furthermore, both estimator and Schwarz Information Criterion methods are also able to estimate the break period without any a priori assumption: recall that in previous work, the break period at 10 days is usually adopted. These two methods not only can be applied for Cepheid studies, as we did in this paper, but also to other astronomical and astrophysical hypothesis testing problems. We also emphasize that our use of the estimator has, for the first time, been generalized to more than two stages and hence is also a statistical result in its own right. In the next section, we outline these techniques in detail and their application to our problems. In Section 3 we apply these methods to LMC Cepheid data and present our results. The conclusions and discussion are given in the last section.

2. THE STATISTICAL METHODS

2.1. The Testimator

The concept of a testimator (or test estimator) was first proposed by Bancroft (1944) in the context of estimating a parameter where a prior guess will be used in place of the estimator of an unknown parameter. The testimator can be applied if the prior guess for the unknown parameter can be ascertained by a test of hypothesis, otherwise the traditional estimator will be used. Due to its superior efficiency compared to traditional estimators, the testimator method has been adapted and refined to suit other situations by Paul (1950), Huntsberger (1955), Bancroft (1964), Arnold & Katti (1972), Bock et al. (1973), Han (1978), Ghosh & Sinha (1988), Yancey et al. (1989), Pandey & Malik (1990), Pandey et al. (1995), Pandey (1997) and Pandey & Srivastava (2001) to name a few. Waikar et al. (1984) and Waikar et al. (2002), in work on two-stage shrinkage estimation, proposed a weighted testimator by placing a weight $1 - k$ on the prior guess and weight k on the traditional estimator, where k is the probability that the guess will be true. They showed that the testimators have far superior efficiency and therefore are more powerful in estimating unknown parameters. This weighted two stage testimation concept can be extended to cover multiple stages. In this paper we apply this “weighted” testimator to investigate the non-linearity of the LMC Cepheid P-L relation as mentioned in the Introduction.

The description of the two-stage testimator method is summarized as follows. For a linear regression of the form of $y = \beta x + a$, the usual least square estimation of the slope to N data points is given as

$$\hat{\beta} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2}, \quad (1)$$

where $\bar{x} = N^{-1} \sum x_i$ and $\bar{y} = N^{-1} \sum y_i$ are the mean

values of x and y , respectively. In the standard hypothesis testing procedure, the null and alternate hypotheses are constructed as $H_0 : \beta = \beta_0$ and $H_a : \beta \neq \beta_0$, respectively, where β_0 is the assumed value of (true) slope given the prior knowledge on the slope. For example, β_0 can be predicted from theoretical calculations. In case that the (true) variance of the slope is known, the z -statistical test (with normal-distribution) can be applied, otherwise the t -statistical test (with T -distribution) will be used for the hypothesis testing. In general the variance is not known, therefore we adopt the t -statistical test in this paper. If the null hypothesis is accepted from the hypothesis testing, the testimator (of the slope), $\hat{\beta}_\omega$, is calculated as (Waikar et al. 1984):

$$\hat{\beta}_\omega = k\hat{\beta} + (1 - k)\beta_0. \quad (2)$$

The constant k in the above equation is defined as

$$\begin{aligned} k &= \frac{|t_{\text{observed}}|}{t_{\text{critical}}}, \\ t_{\text{observed}} &= \frac{\hat{\beta} - \beta_0}{\sqrt{MSE/S_{XX}}}, \\ t_{\text{critical}} &= t_{\alpha/2, \nu}, \end{aligned} \quad (3)$$

where $MSE = (N - 2)^{-1} \sum_{i=1}^N (y_i - \hat{a} - \hat{\beta}x_i)^2$, $S_{XX} = \sum_{i=1}^N (x_i - \bar{x})^2$ and $t_{\alpha/2, \nu}$ is the t -value for $100(1 - \alpha/2)\%$ confidence interval obtained from the associated T -distribution table with $\nu = N - 2$ degree of freedom. Note that the null hypothesis is rejected if $k > 1$. The properties of the testimator are such that:

1. The testimator is an unbiased estimator under H_0 .
2. The testimator has a smaller variance than the usual least square estimator, that is $\text{Var}(\hat{\beta}_\omega) < \text{Var}(\hat{\beta})$.

The proofs for these two properties are given in the Appendix.

2.1.1. Application to the Cepheid P-L Relation

The motivation of this paper is to apply the testimator method to detect any non-linearity in the LMC P-L relation; this has been detected using the F -test (Kanbur & Ngeow 2004; Ngeow et al. 2005). To study any possible variation in slope as the period increases through 10 days, we first sorted the data according to period, from shortest to longest period in $\log(P)$. The sorted sample is then divided into m number of non-overlapping and hence independent subsets according to the Cepheid period. The purpose is to make the bivariate observations independent for each of the subsets. Each of the subsets will then contain n numbers of Cepheids (if the number of data points in the last subset is small, then the last subset will be combined with the previous subset). This enables us to apply the testimator method in multiple stages, together with a conservative Bonferroni testing procedure¹, for detecting such a slope

¹ The Bonferroni testing procedure states that for testing n_g number of hypotheses, the confidence coefficient $(1 - \alpha/2)$ is replaced by $(1 - \alpha/2n_g)$ in each of the hypothesis testings. This is to ensure that the overall confidence coefficient will not be less than the original desired value of $(1 - \alpha/2)$.

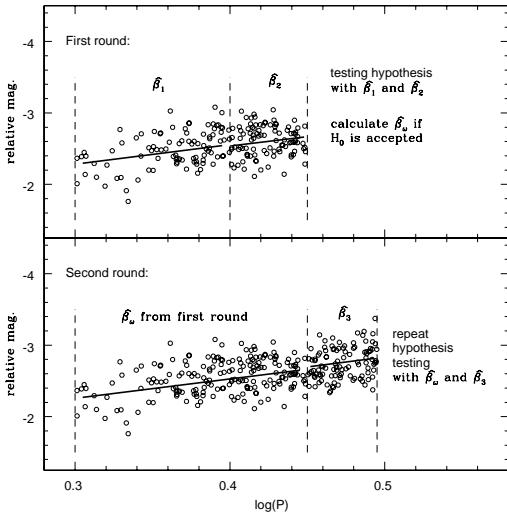


FIG. 1.— Illustration of the estimator procedures. The Cepheid data points are divided to several subsets, sorted according to the $\log(P)$. For the first round, two slopes from the first and second subsets are compared under the hypothesis testing. The estimator, $\hat{\beta}_\omega$, is calculated if the null hypothesis is accepted. In the second round, the estimator from the previous round will be used to compare the estimated slope from the third subset. This is repeated several times until all the subsets have been used up or the null hypothesis is rejected.

variation in the sample. In essence the line of attack is to compute the slope of the first subset and then compare with the slope of the next subset. If the two slopes are “similar”, we look at the slope of the third subset with the smoothed slope obtained from a combination of all the previous subsets. Hence, at the i^{th} round, the slope of a given subset of the data is computed and compared with the smoothed slope from the estimator of all the previous data points. If the two slopes are statistically equivalent, then the current subset of data will be incorporated into the computation of the smoothed slope and compared with the slope of the next subset of data. This smoothness is an important feature since it helps to alleviate, to some extent, the influence of outliers. However, if the slopes are “different”, i.e. a rejection of the null hypothesis, then there is an indication of slope change in the P-L relation. Therefore there will be a total of $n_g = m - 1$ number of hypothesis testings in the multi-stage estimator procedures. In short, the algorithm of applying the estimator method in our case can be summarized as follows:

- In the first round, the slope of first subset, β_1 , is calculated and denoted by $\hat{\beta}_1 = \beta_0$. The slope of the second subset is then compared to β_0 under the null hypothesis of $H_0 : \beta_2 = \beta_0$ and alternate hypothesis of $H_a : \beta_2 \neq \beta_0$. If H_0 is accepted, then the estimator in this round, $\hat{\beta}_\omega^1$, is calculated using equation (2).
- In the second round, the slope of the third subset, β_3 , is calculated and denoted by $\hat{\beta}_3$. The estimator from the first round, represented as $\hat{\beta}_\omega^1 = \beta_0$, is used in the hypothesis testing for this round. The null and alternate hypotheses in this round become $H_0 : \beta_3 = \beta_0$ and $H_a : \beta_3 \neq \beta_0$. If H_0 is accepted,

a new estimator, $\hat{\beta}_\omega^2$, is calculated using equation (2).

- In the i^{th} round, the slope of the $(i + 1)^{th}$ subset, β_{i+1} , estimated by $\hat{\beta}_{i+1}$, is calculated. The estimator from previous $(i - 1)$ round is denoted as $\hat{\beta}_\omega^{i-1} = \beta_0$. The null and alternative hypothesis in this round become $H_0 : \beta_{i+1} = \beta_0$ and $H_a : \beta_{i+1} \neq \beta_0$. If H_0 is accepted, then $\hat{\beta}_\omega^i = k\hat{\beta}_{i+1} + (1 - k)\beta_0$ with k refined from equation (3).
- This is repeated until $i = n_g$ round or the null hypothesis is rejected in the i^{th} round, which indicates a change in slope for the $(i + 1)^{th}$ subset.
- Since in principle there will be a total of n_g hypothesis testings, the Bonferroni testing procedure requires that $t_{critical} = t_{\alpha/2n_g, \nu}$ in each round.

Throughout the paper, we will adopt $\alpha = 0.05$ to ensure the overall confidence level is more than 95% in our test. The first two rounds of our estimator procedures to the study the possible non-linear LMC P-L relation is illustrated in Figure 1.

In order to demonstrate the reliability of this procedure, we apply the estimator method to two simulated data sets: one built from a non-linear P-L relation with a break at 10 days and another one developed from a linear P-L relation. For demonstration purpose, we select one set of the simulated data (out of many simulations) in each cases as representation for testing the estimator method. The plots of these two fake data sets, each of them containing 1500 data points, can be found in figure 1 of Ngeow & Kanbur (2006c). Full details of our procedure for developing these two “fake” data sets are described in Ngeow & Kanbur (2006c). The results of applying the estimator procedures as described to these two fake data sets are given in Table 1. In this table, column 1 denotes the subsets; column 2 gives the range of the period in each subsets; column 3 lists the number of data points, n , in each subsets; column 4 & 5 are the fitted slopes in each subsets and the assigned values of β_0 that used in the hypothesis testing; column 6 & 7 are the observed and critical t -values for each of the hypothesis testing; column 8 & 9 are the corresponding k -value and the outcome of the hypothesis testing; finally column 10 is the values of estimator if the null hypothesis is accepted. Since we know which fake data set is intrinsic linear and non-linear when constructing the P-L relation, we can verify the results found in Table 1. For the fake data with linear P-L relation, our estimator results show that the slopes for each subsets are consistent with the smoothed slopes given from the previous subsets, and the hypothesis testings correctly indicate that there is no changes in slope across all the period ranges. In contrast, the hypothesis testings for the fake data with non-linear P-L relation show that subset 7 has a different slopes than the previous subsets, which indicates a change of slope in this subset. Furthermore, the estimator procedures also correctly pick up the “break period” in subset 7, which brackets the input break period at 10 days, from the outcome of hypothesis testing. There-

TABLE 1
TESTIMATOR RESULTS FOR THE FAKE DATA SETS.

Subset (1)	Period range (2)	n (3)	$\hat{\beta}$ (4)	β_0 (5)	$ t_{\text{observed}} $ (6)	t_{critical} (7)	k (8)	Decision (9)	$\hat{\beta}_\omega$ (10)
"Fake" data set from a linear P-L relation									
1	0.2315-0.4421	200	-2.182 ± 0.403	—	—	—	—	—	—
2	0.4422-0.5005	200	-3.658 ± 0.949	-2.182	1.556	2.718	0.572	accept H_0	-3.027
3	0.5006-0.5508	200	-1.955 ± 1.128	-3.027	0.951	2.718	0.350	accept H_0	-2.652
4	0.5512-0.6079	200	-3.006 ± 1.025	-2.652	0.345	2.718	0.127	accept H_0	-2.697
5	0.6080-0.7349	200	-2.733 ± 0.442	-2.697	0.081	2.718	0.030	accept H_0	-2.698
6	0.7349-0.9610	200	-2.841 ± 0.234	-2.698	0.611	2.718	0.225	accept H_0	-2.730
7	0.9610-1.3553	200	-2.493 ± 0.155	-2.730	1.531	2.718	0.563	accept H_0	-2.597
8	1.3652-2.6170	100	-2.684 ± 0.095	-2.597	0.921	2.748	0.335	accept H_0	-2.626
"Fake" data set from a non-linear P-L relation									
1	0.2315-0.4421	200	-2.442 ± 0.403	—	—	—	—	—	—
2	0.4422-0.5005	200	-3.918 ± 0.949	-2.442	1.556	2.718	0.572	accept H_0	-3.287
3	0.5006-0.5508	200	-2.215 ± 1.128	-3.287	0.951	2.718	0.350	accept H_0	-2.912
4	0.5512-0.6079	200	-3.266 ± 1.025	-2.912	0.345	2.718	0.127	accept H_0	-2.957
5	0.6080-0.7349	200	-2.993 ± 0.442	-2.957	0.081	2.718	0.030	accept H_0	-2.958
6	0.7349-0.9610	200	-3.101 ± 0.234	-2.958	0.611	2.718	0.225	accept H_0	-2.990
7	0.9610-1.3553	200	-2.170 ± 0.155	-2.990	5.281	2.718	1.943	reject H_0	—

NOTE. — See text for the description for each columns. Period ranges are given in $\log(P)$.

fore the estimator method can pick up the P-L relation which is intrinsically non-linear.

2.2. The Schwarz Information Criterion

The problem of deciding whether the LMC Cepheid data are more consistent with two lines of significantly different slopes rather than a single line is exactly analogous to deciding the dimensionality of the model that will fit the given LMC Cepheid data. The method of maximizing the likelihood tends to choose the highest possible dimension. Akaike (1974) suggested maximizing the likelihood subject to a penalty depending on the dimensionality of the model under consideration (Akaike Information Criterion, AIC): $AIC = -2 \ln L + 2k_p$, where L is the likelihood function of the model of dimension k_p (see Takeuchi 2000, as an example in the application of astronomy). However, Schwarz (1978) showed that maximum likelihood estimators can be obtained from large sample limits of Bayes estimates for certain classes of a priori distributions. These distributions only put positive probability on the subspaces of the parameter space corresponding to the competing models. Schwarz (1978) derived the following criterion (Schwarz Information Criterion, SIC; or sometimes also referred as Bayesian Information Criterion, BIC, in the literature): choose the model for which

$$SIC = -2 \ln L + k_p \ln N \quad (4)$$

is a minimum, where N is the total number of data points and $k_p = p + 1$ (with p being the number of fitted parameters, see Schwarz 1978). Some use of the BIC for models selection in astronomical and astrophysical literature can be found, for examples, in Arentoft et al. (2001), Handler et al. (2000, 2002), Koen (1996, 1999, 2006), Koen & Schumann (1999), Koen & Laney (2000), Koen & Lombard (1993, 2003), Liddle (2004, 2007), Mukherjee et al. (1998), Porciani & Norberg (2006) and Sterken et al. (1999).

2.2.1. Application to the Cepheid P-L relation

To test the non-linearity of the Cepheid P-L relation with the SIC method, we consider the models with a

linear P-L relation (the null hypothesis) and a non-linear P-L relation with a break period (in days) at P_0 (the alternate hypothesis) in this paper. For the former case, we have:

$$H_0 : m = \hat{m} = \hat{\beta} \log(P) + \hat{a},$$

$$\text{with } \hat{\sigma}^2 = \frac{1}{N-2} \sum_{i=1}^{i=N} (m_i - \hat{m}_i)^2,$$

$$L = \frac{1}{(\sqrt{2\pi\hat{\sigma}^2})^N} \exp \left[-\frac{1}{2\hat{\sigma}^2} \sum_{i=1}^N (m_i - \hat{m}_i)^2 \right],$$

$$SIC(H_0) = -2 \ln L + 3 \ln N.$$

Similarly, for the alternate model, we have:

$$H_A : m = \hat{m} = \begin{cases} \hat{\beta}_S \log(P) + \hat{a}_S, & \text{for } \log(P) < \log(P_0), \\ \text{with } \hat{\sigma}_S^2 = \frac{1}{N_S-2} \sum_{i=1}^{i=N_S} (m_i - \hat{m}_i)^2, \\ \hat{\beta}_L \log(P) + \hat{a}_L, & \text{for } \log(P) \geq \log(P_0), \\ \text{with } \hat{\sigma}_L^2 = \frac{1}{N_L-2} \sum_{i=1}^{i=N_L} (m_i - \hat{m}_i)^2, \end{cases}$$

$$L = \frac{1}{(\sqrt{2\pi})^N} \frac{1}{(\hat{\sigma}_S)^{N_S}} \frac{1}{(\hat{\sigma}_L)^{N_L}}$$

$$\exp \left[-\frac{1}{2\hat{\sigma}_S^2} \sum_{i=1}^{i=N_S} (m_i - \hat{m}_i)^2 - \frac{1}{2\hat{\sigma}_L^2} \sum_{i=1}^{i=N_L} (m_i - \hat{m}_i)^2 \right],$$

$$SIC(H_A) = -2 \ln L + 5 \ln N.$$

In these expressions, $N_S + N_L = N$ and m is the observed magnitudes after correcting for extinction. The slope (β) and zero-point (a) parameters in the above models are obtained from the maximum likelihood estimation (MLE, which is equivalent to standard least square estimation in our case). Note that the sample variance (σ^2) from MLE is a biased estimate. We corrected the bias with $N_{(L,S)} - 2$ degrees of freedom. For the alternate models, $SIC(H_A)$ is calculated with a range of $\log(P_0)$ that increment in steps of, for example, 0.001. Therefore, a model with linear P-L relation and a range of

models with non-linear P-L relations at different break period are tested with the SIC method. The model with smallest SIC value is the preferred model. In case of $SIC(H_A) < SIC(H_0)$, the minimum value of $SIC(H_A)$ not only suggests that the P-L relation is non-linear, but also gives an estimate of the break period.

To test the SIC method, the same simulated data sets as in the case of the testimator method in Section 2.1.1 were used. For the “fake” data set with linear P-L relation, the values of $SIC(H_0)$ and $SIC(H_A)$ is -164.65 and -161.21 , respectively. While for the “fake” data set with non-linear P-L relation, we found $SIC(H_0) = -100.56$ and the minimum value of $SIC(H_A) = -154.62$ occurs at $\log(P_0) = 0.983$. We then test the SIC method for our application further with various simulations. We first ran two sets of simulations: one set of simulations use the linear P-L relation as input P-L relation, and another set of simulations include the non-linear P-L relation with a break at $\log(P_0) = 1.0$. These simulations mimic the period distribution and the observed dispersions along the P-L relation from the real data. The details for constructing these simulations can be found in Ngeow & Kanbur (2006c). In either sets of the simulations, a large number of simulations is run (typically 1000) and the break period (in $\log[P_0]$) is searched with the SIC method. If the break period cannot be found then this implies the linear P-L relation is the preferred model, and vice versa. The top panels of Figure 2 display the distributions of the break periods from these two sets of data. For the case of a linear P-L relation, the SIC method did not find any break period $\sim 90\%$ of the time. While for the case of no-linear P-L relation, the SIC method detects a range of break period with a peak at $\log(P_0) \sim 1.0$. Therefore, the SIC method can be used to correctly identify the P-L relation that is either intrinsically linear or non-linear at a given break period.

The relatively large dispersion around $\log(P_0) \sim 1.0$ and the long tail toward shorter period that is exhibited in the top panels of Figure 2 can be due to a combination of two effects: (1) the existence of the intrinsic dispersion along the P-L relation; and (2) the non-uniform distribution of the periods in the data (see Ngeow et al. 2005 and Ngeow & Kanbur 2006c for more discussion about the period distribution for Cepheid variables). To portray the impact of these effects on the application of the SIC method for detecting the break period, we ran two additional experiments. The first retains the original period distribution but the intrinsic dispersion of the P-L relation is excluded (however the random photometric errors still persist in the simulation), while the second simulation use a uniform period distribution (in $\log[P]$) and the intrinsic dispersion of the P-L relation is not excluded. The resulted distributions of the break period from SIC method are presented in the bottom panels of Figure 2. It can be seen that the long tail of the distribution present in the top panels is reduced when a uniform period distribution is assumed. Furthermore, if the intrinsic dispersion does not exist in the Cepheid P-L relation, then the SIC method is very efficient to detect the intrinsic break period (at $\log[P_0] = 1.0$ in our case). In reality, the intrinsic dispersion along the P-L relation cannot be eliminated or reduced (at least in the optical bands) and the period distribution of the Cepheid variables will not be uniform (for the reasons discussed in

Ngeow & Kanbur 2006c). We emphasize that the theoretical pulsation modelings are needed to identify the location of the break period or to confirmation the break period at $\log(P_0) \sim 1.0$ (Ngeow et al. 2005).

3. DATA AND RESULTS

In this section, we apply both the testimator and SIC methods to the real LMC Cepheids data in order to investigate whether the V -band Cepheid P-L relation at *mean light* is non-linear or not. We concentrate on the V -band mean light data because the data available in the literature are mostly in the V -band mean light and also because of the evidence for non-linearity as a function of phase is clear (Ngeow & Kanbur 2006a). The data sets we used in this study include the OGLE data adopted from Kanbur & Ngeow (2006) and the MACHO data adopted from Ngeow et al. (2005). They are referred as the “OGLE” sample (with 641 Cepheids) and the “MACHO” data (with 1216 Cepheids), respectively. Note that both data sets have been corrected for extinction using the method described in the corresponding papers. It is also important to point out that these two are independent data sets. To investigate the influence of longer period stars in our testing as well as increasing the number of Cepheids in the OGLE sample, we append the data from Sebo et al. (2002) to the OGLE sample after proper removal of duplication of the Cepheids in both samples and the correction of extinction. This third data set is called “OGLE+SEBO” sample (with 723 Cepheids) and it extends to $\log(P) \sim 2.0$.

The results from using the testimator method to these three LMC Cepheid data sets are summarized in Table 2, with identical layout as Table 1. In the case for the OGLE and OGLE+SEBO data sets, we have tried different sample subset sizes by dividing the samples to $n = 100$ and $n = 150$, which are referred as Test 1 and Test 2 in the table, respectively. In all cases, the testimator method implies that there is a change of slope in the last subset of the samples. Similar results found from Test 1 and Test 2 suggest that our results are not affected by the size of each subset. This indicates the LMC P-L relation becomes non-linear as the period increases through 10 days to longer periods. The last subset also brackets the fiducial break period at/around 10 days: this is consistent with previous work of Ngeow et al. (2005).

The results from using the SIC method are presented in Figure 3 and Table 3 for the same data sets. In Figure 3, the values of SIC for both $SIC(H_0)$ and $SIC(H_A)$ are plotted as a function of the chosen break period, $\log(P_0)$. Since $SIC(H_0)$ is independent of $\log(P_0)$, this represents a straight horizontal line in the figure, and the values of $SIC(H_0)$ for these three data sets are given in Table 3. For the case of $SIC(H_A)$ as a function of $\log(P_0)$, the figure bears witness to the fact that there is a range of $\log(P_0)$ at which the values of $SIC(H_A)$ are smaller than $SIC(H_0)$ in all three data sets. This implies that the non-linear P-L relation is preferred within these period ranges. This result also reinforces the findings of Figure 2 that it is difficult to determine the exact break period of the P-L relation with SIC method (see Section 2.2.1 as well), if it is present. The minimum values for $SIC(H_A)$ found from the figure, and the corresponding $\log(P_0)$ are summarized in Table 3 as well. The confi-

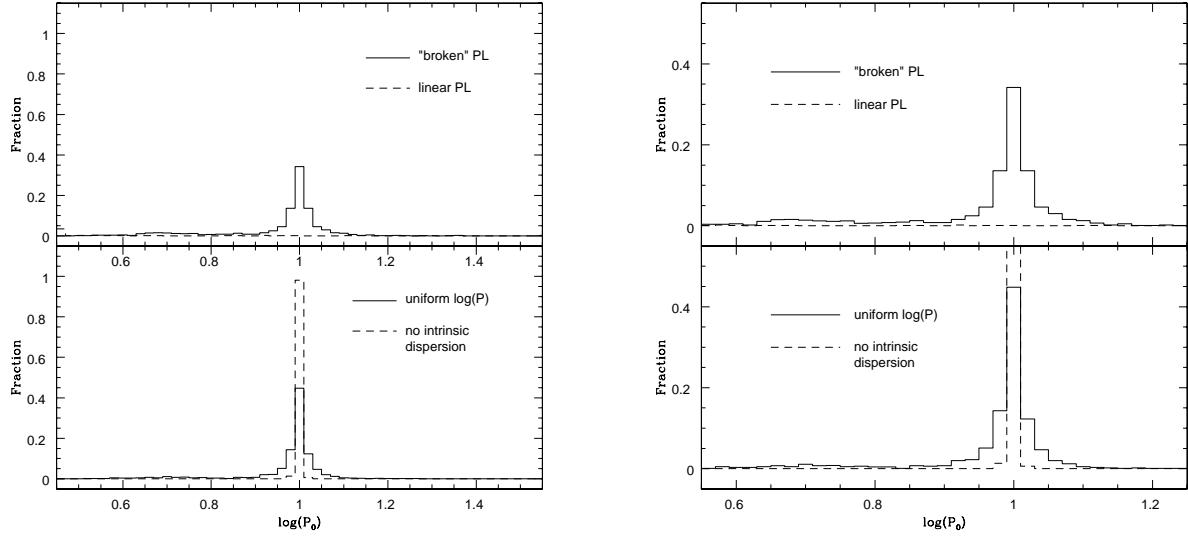


FIG. 2.— Distributions of the estimated break periods, $\log(P_0)$, using the SIC method with various simulations. The top panels show the histograms from two simulations at which the input P-L relations to the simulation is linear (dashed histogram) and non-linear with a break period at $\log(P_0) = 1.0$ (thick histogram), respectively. The bottom panels show the histograms from two additional simulations using the same non-linear P-L relation with a break period again at $\log(P_0) = 1.0$ as input P-L relation: one simulation has uniform distribution of the periods (in $\log[P]$) in the simulated data, and another simulation did not include the intrinsic dispersion of the P-L relation. The right panels show the blown-up region of the left panels.

TABLE 2
TESTIMATOR RESULTS FOR THE REAL DATA SETS.

Subset (1)	Period range (2)	n (3)	$\hat{\beta}$ (4)	β_0 (5)	$ t_{\text{observed}} $ (6)	t_{critical} (7)	k (8)	Decision (9)	$\hat{\beta}_\omega$ (10)
OGLE sample, Test 1									
1	0.4022-0.4771	100	-1.427 ± 0.967	—	—	—	—	—	—
2	0.4787-0.5293	100	-2.273 ± 1.399	-1.427	0.605	2.627	0.230	accept H_0	-1.622
3	0.5294-0.5889	100	-0.746 ± 1.095	-1.622	0.800	2.627	0.304	accept H_0	-1.355
4	0.5891-0.6703	100	-1.887 ± 0.675	-1.355	0.788	2.627	0.300	accept H_0	-1.515
5	0.6704-0.7891	100	-3.055 ± 0.703	-1.515	2.193	2.627	0.835	accept H_0	-2.801
6	0.7900-1.6768	141	-2.462 ± 0.082	-2.801	4.106	2.612	1.572	reject H_0	—
OGLE sample, Test 2									
1	0.4022-0.5043	150	-2.547 ± 0.647	—	—	—	—	—	—
2	0.5043-0.5889	150	-1.783 ± 0.641	-2.547	1.193	2.421	0.493	accept H_0	-2.171
3	0.5891-0.7083	150	-2.347 ± 0.401	-2.171	0.438	2.421	0.181	accept H_0	-2.203
4	0.7103-1.6768	191	-2.590 ± 0.075	-2.203	5.139	2.415	2.128	reject H_0	—
OGLE+SEBO sample, Test 1									
1	0.4022-0.4746	100	-0.989 ± 0.882	—	—	—	—	—	—
2	0.4752-0.5242	100	-2.476 ± 1.202	-0.989	1.237	2.693	0.459	accept H_0	-1.672
3	0.5245-0.5729	100	-4.743 ± 1.339	-1.672	2.292	2.693	0.851	accept H_0	-4.286
4	0.5734-0.6469	100	-2.743 ± 0.907	-4.286	1.701	2.693	0.632	accept H_0	-3.311
5	0.6491-0.7320	100	-2.921 ± 0.933	-3.311	0.418	2.693	0.155	accept H_0	-3.250
6	0.7330-0.9071	100	-3.315 ± 0.400	-3.250	0.162	2.693	0.060	accept H_0	-3.254
7	0.9112-2.1268	123	-2.497 ± 0.089	-3.254	8.535	2.682	3.181	reject H_0	—
OGLE+SEBO, Test 2									
1	0.4022-0.4977	150	-2.545 ± 0.546	—	—	—	—	—	—
2	0.4891-0.5729	150	-2.826 ± 0.706	-2.545	0.398	2.529	0.157	accept H_0	-2.589
3	0.5734-0.6831	150	-2.557 ± 0.432	-2.589	0.073	2.529	0.029	accept H_0	-2.588
4	0.6831-0.9071	150	-3.153 ± 0.253	-2.588	2.234	2.529	0.883	accept H_0	-3.087
5	0.9112-2.1268	123	-2.497 ± 0.089	-3.087	6.651	2.536	2.623	reject H_0	—
MACHO sample									
1	0.4008-0.4715	200	-2.391 ± 0.958	—	—	—	—	—	—
2	0.4719-0.5226	200	-1.843 ± 1.189	-2.391	0.462	2.601	0.178	accept H_0	-2.294
3	0.5231-0.5787	200	-2.623 ± 1.127	-2.294	0.292	2.601	0.112	accept H_0	-2.331
4	0.5795-0.6851	200	-1.851 ± 0.809	-2.331	0.594	2.601	0.228	accept H_0	-2.222
5	0.6588-0.7891	200	-2.948 ± 0.524	-2.222	1.385	2.601	0.533	accept H_0	-2.608
6	0.7910-1.4501	216	-2.123 ± 0.122	-2.608	3.991	2.599	1.536	reject H_0	—

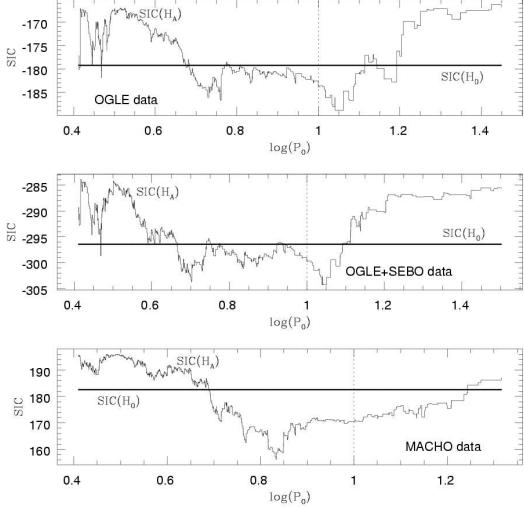


FIG. 3.— Results of the $SIC(H_0)$ and $SIC(H_A)$ as a function of the choosing break period, $\log(P_0)$, for the three LMC Cepheid data sets. The thick horizontal lines are the results for $SIC(H_0)$, which are independent of the chosen break period. The “curves” are the results for $SIC(H_A)$. The horizontal dotted lines represent the chosen break period in the literature (e.g., Tammann & Reindl 2002; Kanbur & Ngeow 2004; Sandage et al. 2004; Ngeow et al. 2005).

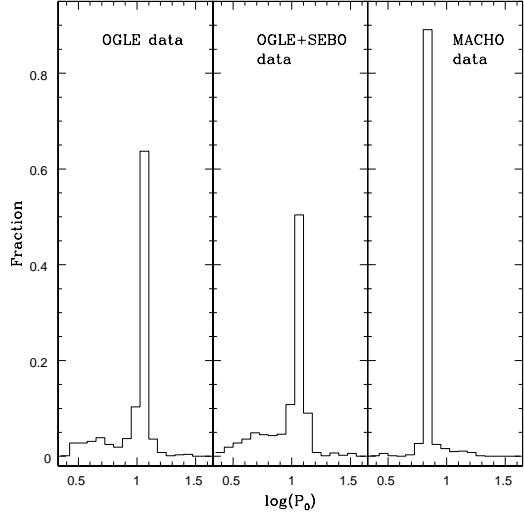


FIG. 4.— Resulted histograms from the bootstrap re-sampling at the break period given in Table 3 for the three LMC Cepheid data sets. See text for details.

dence intervals for the break period can be estimated using bootstrap re-sampling methods. For the model with given $\log(P_0)$ in Table 3, the errors of the regression, $\epsilon_i = m_i - \hat{m}_i$, are randomly drawn (with replacement) to construct a “new” data set, and a new break period is estimated. This is repeated many times to build up the distribution of the break periods. The resulting histograms for these three sets of data are presented in Figure 4. From these distributions, the 5th-, 25th-, 75th- and 95th-percentile are estimated for each of the data sets. The results are given in the last four columns of Table 3. At first glance the break period found from the MACHO data seems to be inconsistent with the OGLE and OGLE+SEBO results. This is due to the difficulty

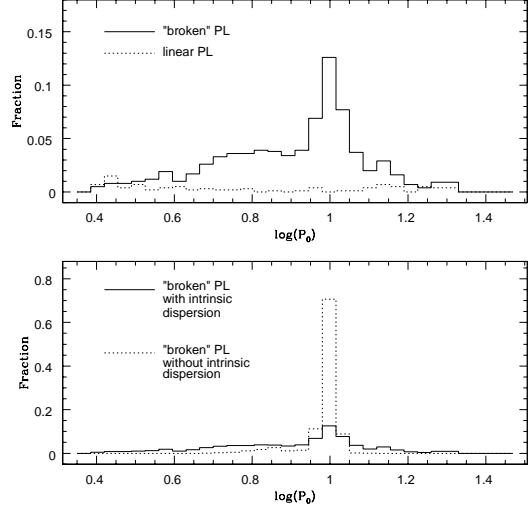


FIG. 5.— Comparisons of the histograms from three sets of simulations for the MACHO data: (1) a simulation that takes a non-linear P-L relation with a break at $\log(P_0) = 1.0$ as input P-L relation and the intrinsic dispersion is included; (2) a simulation with a linear P-L relation as input P-L relation and the intrinsic dispersion is included; and (3) a simulation that takes a non-linear P-L relation with a break at $\log(P_0) = 1.0$ as input P-L relation but without the intrinsic dispersion. Unlike other simulations done in this paper, the periods that go into the simulations are from the *actual* MACHO data.

of accurately estimating the break period with the existence of the instability strip. To demonstrate this, we use the *exact* periods in MACHO data as input periods to our simulations, and generate three different sets of simulations: (1) a simulation with intrinsic non-linear P-L relation; (2) a simulation with linear P-L relation; and (3) a simulation with intrinsic non-linear P-L relation but without the intrinsic dispersion. The resulting histograms for these three sets of simulation are displayed in Figure 5. From this figure it is clear that our result of the break period for MACHO data does not imply an inconsistency to the OGLE and OGLE+SEBO results. The break period found in the data, $\log(P_0) = 0.833$, is within the range of the break periods found from the simulations. This figure also portray the difficulty of estimating the break period from real data when the intrinsic dispersion along the P-L relation is present. Therefore, the break periods given in Table 3 are consistent with the results from testimator (Table 2), the result from non-linear estimation procedure applied in Ngeow et al. (2005, $\log(P_0) = 0.934$ with upper and lower 95% confidence level of 1.089 and 0.778, respectively) and the adopted $\log(P_0) = 1.0$ in the literature. Note that in previous studies (e.g., Tammann & Reindl 2002; Kanbur & Ngeow 2004, 2006; Sandage et al. 2004; Ngeow et al. 2005; Ngeow & Kanbur 2006b) the break period is conveniently chosen to be at 10 days, which is represented as dotted vertical line in Figure 3. The SIC results also supported the non-linear P-L relation to be the preferred model at $\log(P_0) = 1.0$.

4. CONCLUSION AND DISCUSSION

Using two additional statistical approaches, the method of testimators and SIC, to the OGLE, the OGLE+SEBO and the MACHO Cepheids data, we have found strong statistical evidence for a change of slope in

TABLE 3
SIC RESULTS FOR THE REAL DATA SETS.

Dataset	$SIC(H_0)$	$SIC(H_A)$	$\log(P_0)$	5 th -percentile	25 th -percentile	75 th -percentile	95 th -percentile
OGLE sample	-179.29	-188.86	1.041	0.550	1.002	1.041	1.101
OGLE+SEBO sample	-296.43	-304.28	1.041	0.560	0.922	1.052	1.131
MACHO sample	182.61	156.20	0.833	0.806	0.826	0.838	0.936

the Cepheid P-L relation in the LMC at longer period range. These results also strongly support the previous results obtained from the F -test. Therefore, the *observed* LMC P-L relation is non-linear based on these rigorous statistical tests. This implies that either the LMC P-L relation is indeed non-linear or there are some hidden factors in the analysis (see Ngeow & Kanbur 2006c, for more discussion on this). Furthermore, the break periods, or the range of permissible break periods found from this study are consistent with the conveniently chosen break period at 10 days in previous studies. However, our study, both with real and fake data, implies that it is difficult to accurately estimate the break period with both the estimator and SIC methods. This is due to the existence of the intrinsic dispersion along the P-L relation. The confirmation of the break period at/around 10 days has to be done from stellar pulsation modeling studies.

The implications of a non-linear LMC P-L relation on the extra-galactic distance scale and the Hubble constant have been discussed in Ngeow & Kanbur (2005) and Ngeow & Kanbur (2006b) and will not be repeated here. A number of authors, including Spergel et al. (2006), Tegmark et al. (2006), Macri et al. (2006), Olling (2006) and the reference therein, have commented on how an independent estimate of the Hubble constant accurate to 1% can significantly reduce the error bars on Ω , the total density of the universe. Applying the correct form of the Cepheid P-L relation will help in re-

ducing the systematic error of the Hubble constant in the future studies (Ngeow & Kanbur 2006b,c). Over and above this, if the Cepheid P-L relation does indeed have a change of slope at 10 days, it is important to understand this from a stellar pulsation and evolution point of view and investigate fully the ramifications of this new feature (Kanbur & Ngeow 2006; Marconi et al. 2005).

Ngeow & Kanbur (2006c) has investigated various factors that may cause the observed non-linear LMC P-L relation, including the influence of outliers and lack of longer period Cepheids in the sample. However the results from that study suggest that none of these factors are responsible for the observed non-linear LMC P-L relation. We emphasize that the samples we used in the our studies have been cleaned up for obvious outliers. Further, the estimator approach estimates the slope with a variance which is smaller than the standard formula (property 2 stated in Section 2.1.1) is able to minimize the effect of (additional) outliers by smoothing. Regarding the lacks of longer period Cepheids in the sample, we have use the OGLE+SEBO as an expansion sample to the OGLE sample with the increase of period coverage. Both of the samples have shown the same estimator ans SIC results. Therefore, we believe this should not be the cause for the observed non-linear LMC P-L relation.

The authors would like to thank the referee for useful suggestions. This research was supported in part by NASA through the American Astronomical Society's Small Research Grant Program.

APPENDIX

PROOF FOR THE PROPERTIES OF THE TESTIMATOR

We prove the two properties of the estimator as described in Section 2.1 here. To prove that the estimator is an unbiased estimator under H_0 , we note that the estimator from equation (1) is:

$$\hat{\beta}_\omega = k(\hat{\beta} - \beta_0) + \beta_0,$$

where k is defined in equation (3). Therefore the above expression can be re-written as:

$$\hat{\beta}_\omega = \frac{|\hat{\beta} - \beta_0|(\hat{\beta} - \beta_0)}{t_{\alpha/2,\nu}\sqrt{MSE/S_{XX}}} + \beta_0. \quad (\text{A1})$$

This implies that,

$$E(\hat{\beta}_\omega) = \frac{\sqrt{S_{XX}}}{t_{\alpha/2,\nu}} E\left(\frac{1}{\sqrt{MSE}}\right) E[|\hat{\beta} - \beta_0|(\hat{\beta} - \beta_0)] + \beta_0.$$

Since $E(|z|z) = 0$ for variable $z = \hat{\beta} - \beta_0$ with standard normal distribution, and from the above expression, we obtain

$$E(\hat{\beta}_\omega) = \beta_0$$

as desired. The second assertion states that $\text{Var}(\hat{\beta}_\omega) < \text{Var}(\hat{\beta})$. To proof this, we first re-arrange equation (A1) such that:

$$(\hat{\beta}_\omega - \beta_0)^2 = \frac{(\hat{\beta} - \beta_0)^4}{t_{\alpha/2,\nu}^2 MSE} S_{XX}.$$

Assume $\hat{\beta}$ is normally distributed with $N(\beta_0, \sigma_\beta)$ and define $Z = \frac{\hat{\beta} - \beta_0}{\sigma_\beta}$, then Z has a standard normal distribution with $N(0, 1)$. Note that $\sigma_\beta^2 \equiv \text{Var}(\hat{\beta}) = \sigma^2/S_{XX}$, where σ^2 is the variance of the linear regression $y = \beta x + a$, the above expression is reduced to:

$$(\hat{\beta}_\omega - \beta_0)^2 = Z^4 \frac{\sigma^2}{MSE} \frac{1}{t_{\alpha/2, \nu}^2} \frac{\sigma^2}{S_{XX}}.$$

Hence, we have

$$\text{Var}(\hat{\beta}_\omega) = E(Z^4) E\left(\frac{\sigma^2}{MSE}\right) \frac{1}{t_{\alpha/2, \nu}^2} \frac{\sigma^2}{S_{XX}}, \quad (\text{A2})$$

as the last two terms are constants and $\text{Var}(\hat{\beta}_\omega) \equiv E[(\hat{\beta}_\omega - \beta_0)^2]$. For $E(Z^4)$, since the forth moment of a standard normal distribution (the Kurtosis) is 3, then $E(Z^4) = 3$. For $E(\frac{\sigma^2}{MSE})$, we observe that $\frac{\sigma^2}{MSE} = \frac{1}{MSE/\sigma^2} = (N-2)/\sum(\frac{y_i - \hat{a} - \hat{\beta}x_i}{\sigma})^2$. Therefore, $(N-2)MSE/\sigma^2$ is χ^2 distributed with $\nu = N-2$ degree of freedom. It is well-known that if X is χ^2 distributed with ν degree of freedom, then $E(1/X) = 1/(\nu-2)$, hence $E(\frac{\sigma^2}{MSE}) = (N-2)/(N-4)$. Recall that $\sigma^2/S_{XX} = \text{Var}(\hat{\beta})$, then equation (A2) is reduced to:

$$\text{Var}(\hat{\beta}_\omega) = 3 \frac{(N-2)}{N-4} \frac{1}{t_{\alpha/2, \nu}^2} \text{Var}(\hat{\beta}).$$

If $t_{\alpha/2, \nu} > \sqrt{3(N-2)/(N-4)}$, we then have

$$\text{Var}(\hat{\beta}_\omega) < \text{Var}(\hat{\beta})$$

as the assertion states. Due to the Bonferroni testing procedure, condition $t_{\alpha/2n_g, \nu} > \sqrt{3(N-2)/(N-4)}$ is satisfied when $N > 5$ and $\alpha < 0.1$.

REFERENCES

Akaike, H., 1974, IEEE Transactions on Automatic Control, 19, 716

Arentoft, T., Sterken, C. & Handler, G., 2001, MNRAS, 326, 192

Arnold, J. C. & Katti, S. K., 1972, Journal of Multivariate Analysis, 2, 236

Bancroft, T. A., 1944, Annals of Mathematical Statistics, 15, 190

Bancroft, T. A., 1964, Biometrics, 20, 427

Bock, M. E., Yancey, T. A. & Judge, G. G., 1973, Journal of the American Statistical Association, 68, 109

Ghosh, M. & Sinha, B. K., 1988, Journal of Multivariate Analysis, 27, 206

Han, C. P., 1978, Journal of American Statistical Association, 73, 855

Handler, G., et al., 2000, MNRAS, 318, 511

Handler, G., Romero-Colmenero, E. & Montgomery, M. H., 2002, MNRAS, 335, 399

Huntsberger, D. V., 1955, Annals of Mathematical Statistics, 26, 734

Kanbur, S. & Ngeow, C., 2004, MNRAS, 350, 962

Kanbur, S. & Ngeow, C., 2006, MNRAS, 369, 705

Koen, C., 1996, MNRAS, 283, 471

Koen, C., 1999, MNRAS, 309, 769

Koen, C., 2006, MNRAS, 365, 489

Koen, C. & Lombard, F., 1993, MNRAS, 263, 287

Koen, C. & Schumann, R., 1999, MNRAS, 310, 618

Koen, C. & Laney, D., 2000, MNRAS, 311, 636

Koen, C. & Lombard, F., 2003, MNRAS, 343, 241

Liddle, A., 2004, MNRAS, 351, L49

Liddle, A., 2007, astro-ph/0701113

Macri, L., Stanek, K., Bersier, D., Greenhill, L. & Reid, M., 2006, ApJ, 652, 1133

Marconi, M., Musella, I. & Fiorentino, G., 2005, ApJ, 632, 590

Mukherjee, S., Feigelson, E. D., Jogesh Babu, G., Murtagh, F., Fraley, C. & Raftery, A., 1998, ApJ, 508, 314

Ngeow, C. & Kanbur, S., 2005, MNRAS, 360, 1033

Ngeow, C., Kanbur, S., Nikolaev, S., Buonaccorsi, J., Cook, K. & Welch, D., 2005, MNRAS, 363, 831

Ngeow, C. & Kanbur, S., 2006a, MNRAS, 369, 723

Ngeow, C. & Kanbur, S., 2006b, ApJ, 642, L29

Ngeow, C. & Kanbur, S., 2006c, ApJ, 650, 180

Olling, R., 2006, MNRAS, submitted (astro-ph/0607607)

Pandey, B. N. & Malik, H. J., 1990, Communications in Statistics - Theory and Methods, 19, 629

Pandey, B. N., Malik, H. J. & Dubey, P. K., 1995, Communications in Statistics - Theory and Methods, 24, 2261

Pandey, B. N., 1997, Communications in Statistics - Theory and Methods, 26, 2191

Pandey, B. N. & Srivastava, A. K., 2001, IAPQR, 26, 109

Paul, A. E., 1950, Annals of Mathematical Statistics, 21, 539

Porciani, C. & Norberg, P., 2006, MNRAS, 371, 1824

Sandage, A., Tammann, G. A. & Reindl, B., 2004, A&A, 424, 43

Schwarz, G., 1978, Annals of Statistics, 6, 461

Spergel, D., et al., 2006, ApJ, in-press (astro-ph/0603449)

Sterken, C., Broens, E. & Koen, C., 1999, A&A, 342, 167

Takeuchi, T., 2000, Astrophys. & Space Sci., 271, 213

Tammann, G. A. & Reindl, B., 2002, Astrophys. & Space Sci., 280, 165

Tegmark, M., et al., 2006, Phys. Rev. D, 74, 123507

Sebo, K., Rawson, D., Mould, J., Madore, B., Putman, M., Graham, J., Freedman, W., Gibson, B. & Germany, L., 2002, ApJS, 142, 71

Waikar, V. B., Schuurmann, F. J. & Raghunathan, T. E., 1984, Communications in Statistics - Theory and Methods, 13, 1901

Waikar, V. B., Ratnaparkhi, M. V. & Schuurmann, F. J., 2002, *Monte Carlo and quasi-Monte Carlo methods 2000: Proceedings of a conference held at Hong Kong Baptist University, Hong Kong SAR, China*, Ed. Fang, K.-T., Hickernell, F. & Niederreiter, H., Springer, Berlin, pg. 487

Weisberg, S., 1980, *Applied Linear Regression*, 1st Ed., Wiley, New York

Yancey, T. A., Judge, G. G. & Bohrer, R., 1989, Econometrica, 57, 1221